The Impact of Transforming Unstructured Data into Structured Data on a Churn Prediction Model for Loan Customers

Hoon Jung[1] and Bong Gyou Lee[2*]
[1] Hana Institute of Finance, Seoul, 07321, South Korea
[e-mail: rudyhoon@gmail.com]
[2] Yonsei University, Seoul, 03722, South Korea
[e-mail: bglee@yonsei.ac.kr]
* Corresponding Author: Bong Gyou Lee

Received August 24, 2020; revised November 8, 2020; accepted November 27, 2020; published December 31, 2020

Abstract

With various structured data, such as the company size, loan balance, and savings accounts, the voice of customer (VOC), which is text data containing contact history and counseling details was analyzed in this study. To analyze unstructured data, the term frequency–inverse document frequency (TF-IDF) analysis, semantic network analysis, sentiment analysis, and a convolutional neural network (CNN) were implemented. A performance comparison of the models revealed that the predictive model using the CNN provided the best performance with regard to predictive power, followed by the model using the TF-IDF, and then the model using semantic network analysis. In particular, a character-level CNN and a word-level CNN were developed separately, and the character-level CNN exhibited better performance, according to an analysis for the Korean language. Moreover, a systematic selection model for optimal text mining techniques was proposed, suggesting which analytical technique is appropriate for analyzing text data depending on the context. This study also provides evidence that the results of previous studies, indicating that individual customers leave when their loyalty and switching cost are low, are also applicable to corporate customers and suggests that VOC data indicating customers' needs are very effective for predicting their behavior.

***Keywords*:** Churn Prediction Model, Text Mining, Unstructured Data, Voice of Customer, Convolutional Neural Network

This paper is based on the doctoral dissertation of the first author (Hoon Jung, 2020, Exploring the Methods of Transforming Unstructured Data into Structured Data and Their Impact on a Churn Prediction Model, The graduate School of Yonsei University).

http://doi.org/10.3837/tiis.2020.12.005 **ISSN : 1976-7277**

1. Introduction

With the wide popularization of digital devices and the exponential progress of network technology and computing power, the volume of data from heterogeneous sources has increased significantly [1]. These data can be either structured or unstructured, and the amount of unstructured data is increasing more rapidly than the amount of structured data [2, 3]. In spite of these facts, data analysts still mainly depend on structured data when analyzing data in practice [4, 5]. This indicates that there is considerable room for improvement with regard to harnessing unstructured data. Moreover, even though research has been actively conducted on predictive analytics using structured data [6], few studies have been performed to analyze the difference in predictive performance between models based on structured data and models based on both structured and unstructured data. This study evaluates the predictive performance of various forecast models and examines their differences by analyzing corporate customer data of a major bank in South Korea. For understanding the needs of customers, it is important for companies to have an accurate understanding of customer requirements and to develop a correlation between their product outputs and the inputs obtained from their customers [7]. In this sense, analyzing Voice of Customer (VOC) data is critical [8]. VOC refers to what customers communicate to the company such as their complements, complaints, questions, and requirements in the form of text, sound, images, or video. These data are captured in various ways, including direct discussion or interviews, surveys, focus groups, customer specifications, observations, warranty data, and field reports [9]. However, few scholars have empirically examined the impact of using VOC data to detect the intention of customers [10]. The present study investigates how using VOC data improves the predictive performance of forecast models for customer churn. Additionally, by performing various experiments, the optimal technique for transforming the unstructured data into structured data for predictive models is identified.

2. Related Works and Research Hypothesis

2.1 Text as Unstructured Data

Structured data is searchable in a straightforward manner, making it easy to pinpoint information and access it in a fast and fixed way [11]. In contrast, for unstructured data, time-consuming tasks are typically needed to capture specific information and make effective use of it [12]. Unstructured data also include information that does not have a predefined data model and is not organized in a predefined manner [13]. However, when data analysts process the data, there is no formal framework whereby structured and unstructured data can be simultaneously used for gathering business insights to make informed decisions [14]. From a practical viewpoint, there are two pragmatic solutions for combining structured and unstructured data: (1) simply adding the unstructured data to the structured data in a dataset and (2) transposing the unstructured data into the structured data. Text expresses a vast and rich range of information, but it encodes the information in a form that is difficult to decipher automatically [15]. With many types of developed techniques, text mining analyzes the frequency and relative importance of words in documents, identifies semantic links between keywords, and unearths the underlying sentiments, intentions, and attitudes of writers. Moreover, text mining broadens the information base as a starting point for foresight activities and is useful for exploiting the steadily increasing volume of textual data, e.g., from social networking sites and news articles [1, 16]. However, research on

systematic methodologies for determining the optimal text mining approach depending on the features of the text has been insufficient. Grimmer and Stewart [17] proposed an overview of the text as data method but it does not consider the features of the text in detail. Herein, we propose a methodology for converting unstructured data into structured data as well as a process for selecting the optimal analytical method depending on the features of the text.

2.2 The Churn of Corporate Customers

According to many prior studies, customers are "locked in" when their switching costs are high [18, 19]. From a practical standpoint, switching costs can be defined as the perceived economic and psychological costs associated with changing from one alternative to another [20]. However, most studies on customer churn have focused on individual customers; few scholars have examined the churn of corporate customers. In most cases, the business-to-business (B2B) market has a far smaller target audience than the business-to-consumer market. The price may also vary for customers in the B2B market, as customers may agree to place large orders or negotiate special terms. For these reasons, in the case of B2B marketing, most corporate customers are managed through one-to-one marketing rather than mass marketing [21], and close relationships with corporate customers are essential for customer retention [22]. However, many financial companies do not have an analytical model for corporate customer churn, and most companies still rely on speculation based on the experience of salespersons or the customer managers to forecast the customer behavior. Moreover, the average sales and profits for corporate customers are significantly higher than those for individual customers in the financial services industry (According to the data from the bank studied in this paper, the per capita profit of corporate customers is approximately 18 times higher than that of individual customers). This situation supports the claim that financial companies need to construct a churn forecasting model to retain corporate customers.

2.3 Research Hypothesis

Customers leave the company when their loyalty and switching costs are low. Developing switching costs is a common strategy advocated to increase loyalty in industrial markets [23]. According to the study of Behravan and Rahman [24], both customer engagement and customer interaction are positively correlated with the customer retention. Additionally, Danesh et al. [25] proved that switching barriers as well as customer satisfaction has a positive effect on customer retention. By performing the logistic regression for structured data analysis, hypotheses (H1 to H7) are tested in Chapter 3. Hypothesis H8, which is the most important part of this study, is tested in Chapter 4. In this chapter, to investigate the impact of combining structured data and unstructured data on the churn prediction model for corporate loan customers, we analyze unstructured VOC data and develop predictive models utilizing the term frequency–inverse document frequency (TF-IDF) analysis (H8(a)), semantic network analysis (H8(b)), sentiment analysis (H8(c)), and convolutional neural network (CNN) analysis (H8(d)), which is one of the deep learning approaches. **Table 1** presents the hypotheses of this study.

Table 1. Hypotheses of the study

H1	Having a loan with real estate collateral will be negatively related with the churn rate of corporate loan customers.
H2	Having other products except the loan from the same bank will be negatively related with the churn rate of corporate loan customers.
H3	Transaction with the same bank for long time periods (more than 15 years) will be negatively related with the churn rate of corporate loan customers.
H4	Use of automatic withdrawal will be negatively related with the churn rate of corporate loan customers.
H5	Increase in the ratio of loan from other competing banks to total loan in the past year will be positively related with the churn rate of corporate loan customers.
H6	Having a loan from other competing banks prior to churn will be positively related with the churn rate of corporate loan customers.
H7	Getting a new loan from other competing banks in the past year will be positively related with the churn rate of corporate loan customers.
H8 (a)	The predictive performance of churn model for corporate loan customers will be better when unstructured VOC data is transformed into structured data using TF-IDF analysis than that of churn model without unstructured data.
H8 (b)	The predictive performance of churn model for corporate loan customers will be better when unstructured VOC data is transformed into structured data using semantic network analysis than that of churn model without unstructured data.
H8 (c)	The predictive performance of churn model for corporate loan customers will be better when unstructured VOC data is transformed into structured data using sentiment analysis than that of churn model without unstructured data.
H8 (d)	The predictive performance of churn model for corporate loan customers will be better when unstructured VOC data is transformed into structured data using deep learning than that of churn model without unstructured data.

3. Performance of Churn Prediction Model Based on Structured Date

3.1 Data

All the experiments in this study used data from one of the biggest banks in South Korea. The bank has more than 20 million individual customers and approximately 0.3 million corporate customers. Similar to many previous studies on predictive models for customer churn, the dataset used in this study contains information on the profile of customers and their transaction log as structured data. 24 variables were selected as the most relevant data, and these data contains information regarding corporate customers who obtained a loan from the bank. The data used in this study were related to 10,229 companies who had obtained a loan from the bank as of August 2017. The target variable of the predictive model indicates whether a customer paid a loan and in the same month took out a new loan from another competing bank from September 2017 to July 2018. The credit bureau offers this information to all the banks in Korea. Using the SAS 9.4 software package to analyze the data, forecast models employing the variables in the form of structured data were constructed.

3.2 Data Analysis and Results

According to the results of chi-squared tests as shown in **Table 2**, customers who took out a new loan or increased the amount of a loan from another competing bank had a relatively high likelihood of leaving the company. This suggests that some of the signs can be detected before the customer leaves. Furthermore, customers are less likely to leave when they have a real-estate mortgage loan and when they have been trading for a long time (more than 15 years) with the same bank, which indicates that they are “locked in” to the bank. These results are consistent with those of previous studies [18, 19, 26] in which customers were “locked in” when switching costs were high. Consequently, chi-squared tests for hypotheses H1–H7 indicates that the results of the previous studies can be applied to not only individual customers but also corporate customers. The variance inflation factors (VIFs) was calculated

to check whether there was multicollinearity between variables. If VIF is over 10, then there exists multicollinearity. The VIFs of all the variables were less than 4; thus, there was no multicollinearity. A predictive model for customer churn was developed by logistic regression based on the data in **Table 3**. To validate the performance of the predictive model, the data were split into a training set and a validation set in a 7:3 proportion. Two types of measures, the area under the receiver operating characteristic curve (AUC) and Cohen's kappa, were used to evaluate the performance of the predictive model. The AUC is the most widely used statistic for model evaluation [27], and it represents the probability that a randomly chosen positive example is correctly ranked with greater suspicion than a randomly chosen negative example [28]. AUC values are interpreted as follows: a value of >0.9 is excellent, a value between 0.8 and 0.9 is good, a value between 0.7 and 0.8 is fair, and a value of <0.7 is poor [29]. Cohen's kappa, which evaluates the degree of agreement between the classifier and reality [30], is also a good measure for imbalanced data, as it penalizes all-positive or all-negative predictions [31]. Cohen's Kappa ranges from -1 from +1, where +1 represents perfect agreement and -1 represents perfect disagreement between the raters [32]. According to the results, the AUC of the model was 0.72 and the Cohen's kappa was 0.099, indicating that the model performance was not excellent but fair. Hereinafter, customers who leave and remain with the bank are denoted as "churners" and "non-churners," respectively.

Table 2. Hypothesis test results (H1–H7)

Hypothesis	Churn Rate*		Chi-squared and P-value
H1	With real estate collateral	Without real estate collateral	Chi-squared=16.1689, P-value < 0.0001
	1.50%	3.12%	
H2	With other products	Without other products	Chi-squared=4.7583, P-value < 0.0292
	2.26%	2.73%	
H3	Long term transaction	Short term transaction	Chi-squared=18.0038, P-value < 0.0001
	1.91%	2.79%	
H4	Automatic withdrawal	No automatic withdrawal	Chi-squared=31.3661, P-value < 0.0001
	1.80%	2.95%	
H5	Increase in loan from other banks	Decrease or status quo	Chi-squared=30.9911, P-value < 0.0001
	2.78%	1.51%	
H6	Loan from other banks	No loan from other banks	Chi-squared=29.6402 P-value < 0.0001
	2.80%	1.58%	
H7	New loan from other banks	No new loan	Chi-squared=16.1689, P-value < 0.0001
	3.32%	2.24%	

Note : * (Number of corporate customers that paid the entire loan of the bank between September 2017 and July 2018 and took out a new loan from another bank in the same month) ÷ (Number of corporate customers taking out a loan from the bank and having VOC data between June 2017 and July 2018)

Table 3. Descriptions of the variables and the VIFs

Variables	Description
Average Savings	Average of savings account for last 6 month (numerical)
Average Loan	Average of loan balance for last 6 Month (numerical)
Savings Account	Savings account prior to churn (numerical)
Credit Loan	Having credit loan prior to churn (1=yes, 2=no)
Company Type	The size of company (1=small and medium-sized businesses, 2=conglomerate, 3=others)
Deposit	Total deposit prior to churn (numerical)
Number of Employees	The number of employee who works for the company (numerical)
Security Loan	Having security loan prior to churn (1=yes, 2=no)
Foreign Currency	Foreign currency prior to churn (numerical)
Loan Balance	Total loan amount prior to churn (1=less than $ 1.5 million, 2=$1.5 to $8 million, 3=others)
Share of Loan	Change in the ratio of loan to loan from all banks for 1 year (1=decreased less than 10%, 2=increased, 3=decreased more than 10%)
Loan from Other Banks	Number of loans from other competing banks prior to churn (1=1, 2=more than 2, 3=none)
Proximity	Distance between the company and branch of bank (1= within 5 km of the bank, 2=others)
Other Products	The number of products the customer has except loan (1=1, 2=more than 2, 3=none)
Pension	Amount of pension prior to churn (numerical)
Real Estate Loan	Real estate loan prior to churn (1=yes, 2=no)
Security Type	Security type of loan (1=real estate security or other, 2=credit loan, 3=real estate and other)
Change in Loan	The change in total loan for 1 year (1=increased less than $ 0.5 million, 2=increased more than $ 0.5 million, 3=decreased)
New Loan	The number of new loans from other competing banks for 1 year (1=more than 1, 2=none)
Transaction Period	The period since the date of commencing transaction from bank (1=less than 10 years, 2=more than 15 years, 3=10 to 15 years)
Auto-Withdrawal	Amount of automatic withdrawal (e.g. monthly internet fee, 1=less than $ 25 thousand, 2=more than $ 25 thousand, 3=none)
Warranty	Having a guarantee from a government agency (1=Having a guarantee from a government agency, 2=none)
Customer Churn (Dependent Variable)	When customers repay current loans and increase the loan balance (or get a new loan) from other competing banks in the same month (1=churn, 0=no churn)

4. Performance of Churn Prediction Models Based on Structured and Unstructured Data

4.1 VOC Data

In this chapter, one more important variable that was not used in Chapter 3 is considered: the VOC text data written by the customer managers of corporate customers from June 2017 to July 2018. The corporate customer managers of the bank input these data, which include customer complaints, questions and answers, and other communication records between the customer manager and corporate customers, into the VOC system (Some managers input exactly what customers say, and other managers input summarized contents). In most cases, “corporate customer” in this paper refers to the chief executive officer (CEO) of the company that obtains a loan from the bank, but it sometimes refers to the director of the finance department of the company for relatively large companies. Because the VOC data were created by many different customer managers, the text had different styles. The lengths of the text also varied; for example, while some manager wrote only three letters, the longest entry included 2,384 letters in one cell. A total of 610,713 words were used in the data, and the data contained 15,144 nouns when duplicate words were eliminated. To investigate the features of the churners, the VOC data for the churners and the non-churners are analyzed separately. The most frequently occurring words in the VOC data for the churners were “transaction,” “lending,” “loan,” “request,” and “repayment.” In contrast, the most

frequently occurring words in the VOC data for the non-churners were "transaction," "lending," "request," "plan," and "the bank," in order of frequency.

4.2 TF-IDF Analysis

TF-IDF is an analytical technique that calculates the frequency of words appearing within documents for deriving the relative importance of each word. The TF-IDF uses the term frequency (TF) and inverse document frequency (IDF) and is effective for excluding subjective interpretations in the case of text. The TF-IDF is calculated by multiplying the TF by the IDF. The TF-IDF has been modified by many researchers [33 -36]; hence, there are various formulas. To find the significant words related to customer churn in the VOC data, the VOC text for churners and non-churners were analyzed separately.

This was done to identify the difference between the words in the VOC data for the churner and non-churner groups, rather than compare the VOC data within the same group. As mentioned previously, the TF-IDF technique identifies many words in each document and measures the similarity between documents; therefore, it is difficult to use the value of the TF-IDF in the case of a paired comparison between the two corpuses of this study. For instance, if a particular word appears in both corpuses, the TF-IDF is 0 regardless of the frequency of the word in each corpus, which can lead to an incorrect interpretation. Hence, the formula considered here for IDF is a logarithmic function with 1 added, as applied in previous studies [35, 37]. The modified formula is as follows:

$$\mathrm{TF}_{i,j} \times \mathrm{IDF}_j = (n_{i,j} \div \sum_k n_{k,j}) \times (\log \frac{|D|}{|\{d_j | t_j \in d_j\}|} + 1)$$

where $n_{i,j}$ represents the number of occurrences of word t_i in a document d_j, $\sum_k n_{k,j}$ represents the total number of words in a document d_j and D represents the number of documents in a group of documents. $\{d_j | t_j \in d_j\}$ represents the number of documents containing the word t_j.

To determine which words occurred more frequently in the VOC data for the churners than in those for the non-churners, the TF-IDF for all words was calculated, separating the VOC into the text for the churners and non-churners. For each word, the TF-IDF of the words churners used was divided by the TF-IDF of the words non-churners used to identify the words that were used more often by churners than by non-churners. This value is denoted as the "Relative Index" of the TF-IDF hereinafter. Interestingly, the word having the highest Relative Index of the TF-IDF was "churn". This means that "churn" was relatively the most often used word in the VOC of churners, whereas it was less frequently used in the VOC of non-churners. We took the 10 top words with regard to the Relative Index of the TF-IDF, and if at least one of these 10 words appeared in the VOC of a customer, the value of the variable "Sign of Churn in Text" was set to 1; otherwise, it was set to 0. Then, a predictive model was built by logistic regression based on the independent variables employed in Chapter 3 and the variable "Sign of Churn in Text." The AUC of the predictive model for which unstructured VOC data were transformed into structured data using the TF-IDF was 0.735591. This is better than the baseline performance (0.720001), the AUC of the predictive model using only structured data, as presented in Chapter 3.

To cross-check the result, Cohen's kappa analysis was also conducted. The Cohen's kappa was 0.1092 when the top 10 words with regard to the TF-IDF were applied to the logistic regression. This value confirms that the predictive performance of the model with the VOC data transformed into structured data via the TF-IDF is better than that of the original model,

because the Cohen's kappa of the benchmark model was 0.099, as reported in Chapter 3. To find a better AUC, the same procedure was conducted with the top 20 words according to the Relative Index of the TF-IDF. If at least one of these 20 words appeared in the VOC of a customer, the value of the variable "Sign of Churn in Text" was set to 1; otherwise, it was set to 0. In the stepwise selection for logistic regression, the "Sign of Churn in Text" was selected again. The AUC was 0.741594 and the Cohen's kappa was 0.1178, indicating that the predictive performance of the model with the top 20 words according to the Relative Index of the TF-IDF is better than that of the model with the top 10 words according to the Relative Index of the TF-IDF. Here, the same procedure was repeated with the top 30 words, the top 40 words, and the top 50 words with regard to the Relative Index of the TF-IDF. A performance comparison of the models based on the AUC is presented in **Table 4**. The results suggest that the AUC was maximized for the top 30 words, indicating that a good cutoff is somewhere between the top 20 words and the top 40 words, as the AUC decreased when the top 40 words were applied. These are heuristics that can vary depending on the features of the data and the variables selected. Therefore, there may be another optimal value of the AUC for enhancing the predictive performance of the model. However, the objective of this study was not to identify the optimal cutoff; thus, it was not necessary to perform additional experiments. As mentioned previously, the AUC of the benchmark model developed using only structured data in Chapter 3 was 0.720001, and its Cohen's kappa was 0.099. The results for both the AUC and Cohen's kappa indicated that the predictive power was better than that of the benchmark model for all cases where the VOC (unstructured data) was converted into structured data using the Relative Index of the TF-IDF. Hence, the foregoing experimental results indicate that the predictive performance was improved by transforming unstructured data into structured data based on the TF-IDF and are consistent with H8 (a).

Table 4. Comparison of the predictive performance between the models

Model		AUC	Cohen's kappa
Benchmark model		0.720001	0.0990
Relative Index of Customer Churn based on Relative Index of TF-IDF	(1) Top 10 words	0.735591	0.1092
	(2) Top 20 words	0.741594	0.1178
	(3) Top 30 words	0.756723	0.1351
	(4) Top 40 words	0.752336	0.0997
	(4) Top 50 words	0.749472	0.1004

4.3 Semantic Network Analysis

A semantic network is a directed graph consisting of nodes that represent domain objects and links that represent semantic relations between them [38]. From a relational perspective, a semantic network is not an independent object; rather it comprises the relationships between the objects, which allow a concept or situation to be understood more clearly. Describing the structured results from uncategorized data rather than fully categorized data, semantic network identifies the relationships between concepts by providing a spatial representation of the language structure and visualize the relationships between the main concepts and other concepts presented in the text [39]. In this study, word co-occurrence analysis, a technique for analyzing word pairs that appear in the same sentence or same document [40], was

performed to identify the important words related to customer churn from the VOC data by using the software package Python 3.8. According to the analysis results, semantic networks consisting of the words appearing in the VOC text are constructed. The procedures of the semantic network analysis are as follows. To investigate the keywords of churners and non-churners separately, the data were divided into two groups consisting of the extracted words from the VOC of churners and the extracted words from the VOC of non-churners. Next, the degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality for the semantic networks of each group were calculated. Using the centrality values for the two groups, we calculated the centrality ratios by dividing the centrality for churners by the centrality for non-churners. The text was written in Korean; hence, semantic network analysis was performed in the Korean language to investigate the original intention of the writers. The software package Gephi 0.9.2 was utilized for the semantic network analysis. To choose the optimal centrality for this study, the predictive performances of the models based on the degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality are compared. Here, we identified the top 10 words with regard to the centrality ratio based on the degree centrality and created the variable "Sign of Churn in Text," which was 1 if these words appeared in the VOC of a customer, and 0 otherwise. By applying the variable which was transformed into the structured data from the VOC texts based on the degree centrality, a logistic regression model for predicting customer churn was developed. The data were split into a training set and a validation set in a 7:3 proportion. By performing stepwise selection for the training set, 10 variables including "Sign of Churn in Text" were selected. The formula derived from the training set was applied to the validation set to calculate the AUC of the predictive model. For the validation set, the AUC of the model was 0.723256, which was better than the AUC of the benchmark model presented in Chapter 3. The Cohen's kappa was 0.0995, which was also better than that of the benchmark model. The same procedure was repeated for the closeness, betweenness, and eigenvector centralities. According to the results shown in **Table 5**, eigenvector centrality-based model outperformed the other models with regard to predictive performance, as it had the highest AUC and Cohen's kappa values. Therefore, the eigenvector centrality was employed as the centrality measure in the following semantic network analysis.

Table 5. Comparison of the predictive performance of the models

Model	AUC	Cohen's kappa
Benchmark model	0.720001	0.0990
Degree Centrality	0.723256	0.0995
Closeness Centrality	0.723960	0.0964
Betweenness Centrality	0.722677	0.0971
Eigenvector Centrality	0.726208	0.0995

Table 6. Comparison of the predictive performance of the models

Model		AUC	Cohen's kappa
Benchmark model		0.720001	0.0990
Eigenvector centrality based	top 10 words	0.726208	0.0995
	top 20 words	0.728588	0.1125
	top 30 words	0.727415	0.1062
	top 40 words	0.727474	0.0971

To find an appropriate cutoff, another predictive model was constructed by applying the top 20 words based on the eigenvector centrality; i.e., if one of the top 20 words appeared in the VOC for a customer, the value was 1; otherwise, the value was 0. Thereafter, the same procedure was repeated for the top 30 words and the top 40 words. These methods are also heuristics to search for a better value and are identical to the procedure employed for the TF-IDF analysis. The data were split into a training set and a validation set in a 7:3 proportion. The stepwise selection and the analysis of the estimate were conducted for the training set, and the AUC and Cohen's kappa were calculated using the data of the validation set. A comparison of the performance of each model in terms of the AUC and Cohen's kappa is summarized in **Table 6**. The results indicated that the predictive performance of the model was the best for both the AUC and the Cohen's kappa when the VOC data were transformed into the structured data with the top 20 words. These findings suggest that the predictive model has the highest performance when the top 20 words based on the eigenvector centrality are applied to the model by the heuristics. Consequently, these results are consistent with H8 (b).

4.4 Sentiment Analysis

Sentiment analysis is a computational study of opinions, sentiments, emotions, and attitudes toward an entity that are expressed in text [41]. In general, sentiment analysis investigates the sentimental polarity or strength values of words using a predefined sentiment dictionary to derive the sensitivity of the entire document [42, 43]. Hence, a sentiment dictionary was needed for the sentiment analysis, and the National University Sentiment Dictionary[1] was used in this study. The sentiment score, i.e., the polarity of words, ranged from –2 to +2 in this study. In the sentiment dictionary, a word with a polarity of +2 was very positive, and a word with a polarity of –2 was very negative. As done for the TF-IDF and the semantic network analysis, the VOC data were transformed into structured data to build a predictive model. For this, the VOC data were converted into a binary variable: the value of "Sign of Churn in Text" was 1 if the sentiment score was –2, and was 0 otherwise. However, the variable "Sign of Churn in Text" was not selected as an independent variable in the stepwise selection.

To verify the result, a logistic regression was performed without the stepwise selection. "Sign of Churn in Text" had no statistically significant effect on the dependent variable (Pr > Chi-sq: 0.1359). The VOC data were converted into a binary variable again for re-examination; here, the value of "Sign of Churn in Text" was 1 if the sentiment score was –2 or –1, and it was 0 otherwise. However, this variable was not selected in the stepwise selection, yielding the same result. Moreover, this "Sign of Churn in Text" variable had no statistically significant effect on the dependent variable (Pr > Chi-sq: 0.175) according to the results of logistic regression without the stepwise selection. Finally, the sentiment score was considered as a numerical independent variable without transforming it into to a binary variable. However, again, the sentiment score was not selected in the stepwise selection, and this variable had no statistically significant effect on the dependent variable in the model, according to the results of logistic regression without the stepwise selection. These results suggest that sentiment analysis is not an effective technique for analyzing the VOC data in this study. As mentioned previously, because sentiment analysis is a computational study of

[1] https://github.com/park1200656/KnuSentiLex

sentiments and emotions expressed in text, it may not be suitable for business documents. The customer managers of the bank generally did not write about sentiments and emotions; thus, sentiment analysis was not useful for analyzing the text data in the VOC in this study. These findings do not support H8(c).

4.5 Convolutional Neural Network Analysis

Deep learning has recently attracted considerable interest owing to the advent of efficient parallel solvers optimized for modern graphics processing units [44]. One type of deep learning model, the CNN utilizes layers with convolving filters that are applied to local features [45, 46]. CNN models are effective for natural language processing and have achieved excellent results in semantic parsing and sentence modeling [46]. In the present study, the data were split into 55% for a training set, 15% for a validation set, and 30% for a test set to develop and test CNN models. The letters in the training set were replaced with corresponding numbers based on the UTF-8 code. Next, word embedding process was performed to map the numbers to a vector of 128 dimensions with a length of 1000 using Keras, which is a Python library for machine learning. Word embedding is used to represent the words in a continuous and multidimensional vector space, so that it is easy to capture the semantic similarity between words by calculating the vector distance [47]. For the activation function, we employed the rectified linear unit (ReLU), i.e., $f(x) = 0$ for $x < 0$ and $f(x) = x$ for $x \geq 0$, which is one of the most widely used activation functions for deep neural networks [48, 49], instead of the sigmoid function. When the sigmoid activation function is used, the "vanishing gradient problem" can occur, which means that the gradient becomes vanishingly small with recurrent multiplication to compute the gradients of the other layers as the sigmoid function has gradients in the range of 0 to 1.

The CNN model consisted of a convolution layer, max pooling layer, and fully connected layer [50]. The size of the CNN filter was 2×128, 3×128, 4×128, and 5×128 in the convolution layer of this study, which means that the filters identify the characteristics of two consecutive letters, three consecutive letters, four consecutive letters, and five consecutive letters, respectively. To identify 64 characteristics for each consecutive letter, 64 of these filters were employed in the present study. Here, the score for predicting the customer churn developed by the character-level CNN model was applied to the logistic regression as an independent variable. This can be considered as "a model in a model," as the logistic regression is also a predictive model. The CNN score, which was based on CNN analysis, was selected the independent variable of the model. It had a positive relation with the churn rate of the corporate customer, considering that the estimate was 3.9341 and the Pr > Chi-sq was less than 0.0001. The result indicates that the AUC was 0.864919, representing a significant improvement compared with that of the benchmark model. Additionally, the Cohen's kappa indicated that the predictive performance of this model was superior to that of the other models. Both a character-level CNN model and a word-level CNN model were developed in the present study; thus, the aforementioned procedures were conducted twice for each CNN model. Taking the word "loan" as an example, the letters of the word ("l," "o," "a," "n") are each considered as an independent unit when analyzing the data via the character-level CNN. However, for the word-level CNN, the word "loan" is considered as an independent unit in the analysis. According to the results of applying the predictive score obtained using the word-level CNN, the AUC of the model was 0.845907, and the Cohen's kappa was 0.2241. Thus, the character-level CNN model outperformed the word-level CNN model in this study. With regard to the predictive performance, both the character-level and word-level CNN models outperformed the benchmark model developed

without unstructured data, as presented in Chapter 3. These findings are consistent with H8(d). Interestingly, in previous studies involving comparisons between predictive models for English text, character-level CNNs outperformed word-level CNNs. Zhang et al. (2015) compared the performance of a character-level CNN, a word-level CNN, n-grams and their TF-IDF variants, and a bag-of-words and demonstrated that the character-level CNN outperformed the others for English text. Similarly, Vijayaraghavan et al. [51] performed a social-media analysis and found that a character-level CNN model outperformed a word-level CNN model for English text. Furthermore, Kim et al. [52] reported that a character-level CNN outperformed a word-level LSTM for English, Arabic, Czech, French, German, Spanish, and Russian. In this context, as it has been demonstrated in the present study that a character-level CNN outperformed a word-level CNN for Korean, the argument that character-level deep learning generally outperforms word-level deep learning in many languages is supported by the results of this study.

Table 7. Comparison of the performance of the models

Method		AUC	Cohen's kappa
The benchmark model		0.720001	0.0990
Model based on structured & unstructured data	TF-IDF	0.756723	0.1351
	Semantic Network	0.728588	0.1125
	Sentiment Analysis	N/A	N/A
	character-level CNN	0.864919	0.2574

5. Comparison of Performance of Predictive Models and Selection of Optimal Technique

5.1 Comparison of Performance of Models

In this study, many sets of experiments were performed in various ways. A comparison of the predictive performance of the models based on the AUC and Cohen's kappa is presented in **Table 7**. The results indicate that the predictive model for which the VOC unstructured data were transformed into structured data, outperformed the benchmark model. Further details of the results of these experiments are as follows. Regarding the AUC, the logistic regression model using the score derived by the character-level CNN exhibited the best performance (0.864919), followed by the model using the score derived by the word-level CNN (0.845907), the model using the variable based on the TF-IDF (0.756723), and the model using the variable obtained via semantic network analysis (0.728588). The AUC of the benchmark model developed with only the structured data as described in Chapter 3 was 0.720001. Regarding the Cohen's kappa, the logistic regression model using the score derived by the character-level CNN also exhibited the best performance (0.2574), followed by the model using the score derived by the word-level CNN (0.2241), the model using the variable based on the TF-IDF (0.1351), and the model using the variable obtained via semantic network analysis (0.1125). The Cohen's kappa of the benchmark model developed with only the structured data was 0.0090. The order of the model performance was identical regardless of the evaluation criteria (AUC and Cohen's kappa). These results indicate that transforming the VOC unstructured data into structured data significantly improved the predictive performance of the churn forecasting model. Moreover, the results revealed that the predictive model using VOC data outperformed the benchmark model, indicating that the

VOC is one of the key drivers to forecast customer behavior. In contrast, the model using the variable derived from sentiment analysis did not exhibit significant predictive performance, as explained in Chapter 4.

5.2 Optimal Technique and Methodology

Various analytical techniques were used in this study, such as logistic regression; TF-IDF analysis; semantic network analysis; a character-level CNN; and a word-level CNN. According to the results of these experiments, a methodology is proposed to identify the optimal technique depending on the features of the text and the analysis objective (**see Fig. 1**). First, if it is important to identify the meaning of the text, sentiment analysis or semantic network analysis is recommended. In this case, if there are numerous emotional expressions in the text, sentiment analysis is recommended; otherwise, semantic network analysis should be considered. Second, if it is not important to identify the meaning of the text, TF-IDF or a deep learning technique such as CNN is recommended. In this case, if the document is long and its style is varied, a CNN is recommended; otherwise, the TF-IDF can be considered. As discussed in Chapter 4, if data analysts wish to apply unstructured data as independent variables to predictive models, the data normally must be transformed into a structured form, because analytical models require structured data in most cases. Text data can be transformed into binary, nominal, categorical, and numerical values. Moreover, by using a sentiment score or predictive score, a data analyst can transform text data into structured data and analyze documents. These scores can be considered as independent variables or can be multiplied by the frequency of occurrence of each word in the text. The detailed method depends on the performance of each case, and the optimal method can be selected based on the predictive power of the model.

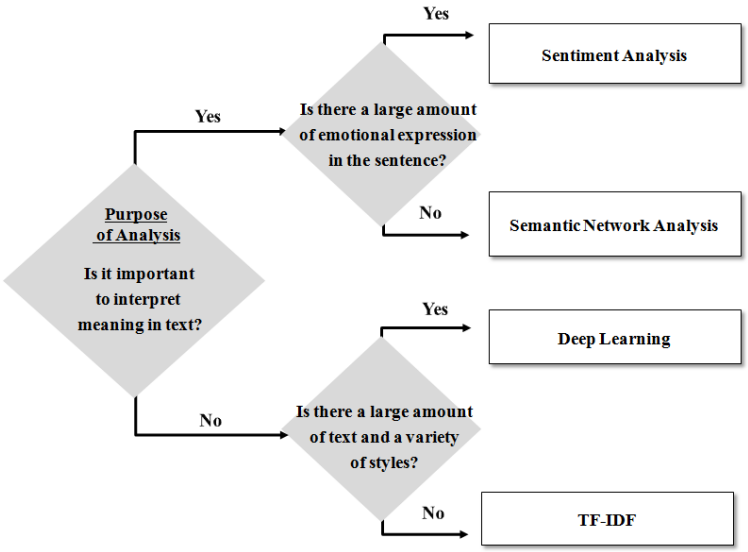

Fig. 1. Model for selection of the text mining technique

6. Conclusions

6.1 Theoretical Contributions

The various experiments of the present study have demonstrated that the VOC data enhanced

the predictive performance of the model for customer churn. Furthermore, this study is meaningful in that it targeted corporate customers, who have been studied less frequently than individual customers in previous studies on forecast models. To build the predictive model for customer churn efficiently and to improve its accuracy, the VOC data were transformed into structured data. We also investigated various methods for combining structured data and unstructured data and evaluated the degree of improvement in the predictive performance for each model via different analysis techniques, such as the TF-IDF analysis, semantic network analysis, sentiment analysis, and CNN analysis. The results indicated that the performance of a deep learning technique, CNN was the best, as the size of the VOC data was very large and the text was written in various styles by multiple managers. In particular, the character-level CNN model outperformed the word-level CNN model with regard to the predictive power, which is similar to the results of previous studies for English and other languages. Additionally, the analysis results indicate that sentiment analysis is not suitable for business documents, as they normally have limited emotional expression. Based on the experimental results, a selection model was proposed to determine the optimal analytical technique, depending on the purpose of the text mining, the features of the document, and the writing styles. With regard to the analysis purpose, if it is important to identify the meaning or intention in the document, sentiment analysis and semantic network analysis are recommended. In this case, sentiment analysis should be considered if there are numerous emotional expressions in the document; otherwise, semantic network analysis is recommended. In contrast, if it is not important to identify the meaning or intention in the document, deep learning techniques such as a CNN and word frequency-based analysis such as the TF-IDF are recommended. In this case, deep learning should be considered if the document is large and diverse in style; otherwise, TF-IDF analysis is recommended. The proposed model for selecting analytical techniques provides specific guidelines to help data analysts choose the optimal method for analyzing text data. Moreover, the reasons why corporate customers leave the company were explored, focusing on the loan business of the banking industry. Many studies have been performed on customer churn; however, most of them focused on individual customers, even though corporate customers are very profitable for the bank [53]. In this study, the raw data of corporate customers from a bank was analyzed, and it was found that corporate loan customers who have mortgage loans are less likely to leave. Additionally, it was also found that corporate customers who use an automatic withdrawal service are less likely to leave the bank. These findings are consistent with the reasoning that individual customers with a high switching cost cannot easily leave the company.

6.2 Managerial Implication and Future Research

According to results presented in Chapter 3, having other products besides the loan from the bank is negatively related to the churn rate of corporate loan customers. Therefore, it is important to increase the number of products customers have by actively cross-selling other products of the company, such as savings accounts. This will not only increase profitability but also help to retain customers. Furthermore, because transaction with the same bank for long time periods is negatively correlated with the churn rate of customers, it is important to retain customers as well as attract new customers. Hence, it is necessary to establish a systematic incentive system for long-term customers. Lastly, obtaining a loan from other competing banks was positively correlated with the churn rate of customers. Because financial companies can acquire this information from credit bureau records, it is important to monitor signs of churn by analyzing customer behavior in advance. For more effective

processing of unstructured data, the methodology for integrating unstructured and structured data in the dataset was addressed in this study. The text can be transformed into a binary variable, categorical variable, or numerical variable according to the data analysts' decision to improve the predictive performance of the model. Using various analysis techniques, we detected the important words used in the text of churners and transformed them into binary variables indicating whether the keywords were used in the documents of customers. Data analysts can also transform text data into categorical variable or numerical variables according to the frequency of occurrence of the keywords in the documents of customers. In Chapter 4, a methodology was presented for developing a predictive model utilizing the score based on another analytical model, such as deep learning. Thus, the logistic regression model for forecasting customer churn used the predictive score based on a CNN as an independent variable. Here, both character-level CNN and word-level CNN models for unstructured VOC data were developed, and the unstructured data were transformed into structured data by using the scores of CNN models, after which another predictive model for structured data was constructed. The findings in Chapter 4 indicate that the VOC is very useful for predicting the churn of customers; therefore, the data must be managed systematically. However, customer managers tend to neglect this valuable information because it is not directly applicable in a traditional marketing context and there is seldom in-house knowledge on how to convert VOC data into an analyzable form [14]. Moreover, few companies have a VOC system with an interface allowing customer managers to input and monitor VOC data. As discussed previously, because long-term relationships with corporate customers are very important, VOC data must be managed consistently, even when customer managers are replaced or leave the company. Furthermore, quality management of the VOC system is important, as we observed some poor data inputs and cases of insufficient information in the VOC data of this study. Therefore, companies must build a VOC system, monitor the VOC data, and set specific guidelines for inputting VOC data. There is a nearly universal consensus that the main drawback of deep learning technique is the non-interpretability, which refers to the lack of interpretability of the features in the model [54, 55], although there have been recent efforts toward mitigating this drawback [56]. Therefore, even though the CNN models, particularly the character-level CNN, exhibited excellent performance in this study, it is difficult to determine why a customer leaves, by using the CNN predictive model. Of course, financial companies create predictive models for customer churn not only to identify customers' churn scores but also to prevent them from leaving the company by understanding the factors that affect customer churn. Therefore, there is still a need to utilize classical data mining techniques such as logistic regression to predict customer behavior, rather than simply relying on deep learning methods. Additionally, the text data analyzed in this study had limitations for detecting the detailed intentions and emotions of customers, as the VOC data were written by corporate customer managers of the bank. Analyzing Speech-To-Text (STT) data obtained by converting a recording of a customer's voice from a call center into text data may yield different results, particularly in sentiment analysis. Hence, even though the results of the sentiment analysis were not statistically significant in this study, if VOC data are analyzed in the STT form, the result of the predictive model with sentiment analysis can be closely related to the customer churn.

References

[1] V. Kayser and K. Blind, "Extending the knowledge base of foresight: The contribution of text mining," *Technological Forecasting and Social Change*, vol. 116, pp. 208-215, Mar. 2017. Article (CrossRef Link)

[2] T. K. Das and P. M. Kumar, "Big Data Analytics: A Framework for Unstructured data Analysis," *International Journal of Engineering and Technology*, vol. 5, no. 1, pp. 153-156, Mar. 2013. Article (CrossRef Link)

[3] Q. He, C. A. W. Glas, M. Kosinski, D. J. Stillwell, and P. B. Veldkamp, "Predicting self-monitoring skills using textual posts on Facebook," *Computers in Human Behavior*, vol. 33, pp. 69-78, Apr. 2014. Article (CrossRef Link)

[4] C. Hänig, M. Schierle, and D.Trabold, "Comparison of Structured vs. Unstructured Data for Industrial Quality Analysis," in *Proc. of the World Congress on Engineering and Computer Science*, vol. 103, pp. 257-270, July 2011. Article (CrossRef Link)

[5] W. Ji, R. Chen, F. Li, and Q. Ling, "Log Prediction of Wireless Telecommunication Systems Based on a Sequence-To-Sequence Model," *Journal of Advances in Mathematics and Computer Science*, vol. 24, no. 1, pp. 1-8, Aug. 2017. Article (CrossRef Link)

[6] L. Dey, H. Meisheri, and I. Verma, "Predictive Analytics with Structured and Unstructured Data - A Deep Learning Based Approach," *IEEE Intelligent Informatics Bulletin*, vol. 18, no. 2, pp. 27-34, 2017. Article (CrossRef Link)

[7] C. C. Aguwa, L. Monplaisir, and O. Turgut, "Voice of the customer: Customer satisfaction ratio based analysis," *Expert Systems with Applications*, vol. 39, no. 11, pp. 10112-10119, Sep. 2012. Article (CrossRef Link)

[8] F. Fatahillah, B. N. Saryanto, and E. Rimawan, "Chartering Services Development with the QFD Approach: Case Study on Liquid Freight Shipping Companies," *International Journal of Innovative Science and Research Technology*, vol. 4, pp. 457-464, 2019. Article (CrossRef Link)

[9] A. S. Khangura and S. K. Gandhi, "Design and Development of the Refrigerator with Quality Function Deployment Concept," *International Journal on Emerging Technologies*, vol. 3, no. 1, pp. 173-177, Apr. 2012. Article (CrossRef Link)

[10] P. Li, Y. Yan, C. Wang, Z. Ren, P. Cong, H. Wang, and J. Feng, "Customer Voice Sensor: A Comprehensive Opinion Mining System for Call Center Conversation," in *Proc. of IEEE International Conference on Cloud Computing and Big Data Analysis*, pp. 324-329, 2016. Article (CrossRef Link)

[11] M. Gärtner, A. Rauber, and H. Berger, "Bridging structured and unstructured data via hybrid semantic search and interactive ontology-enhanced query formulation," *Knowledge and Information Systems*, vol. 41, pp. 761-792, 2014. Article (CrossRef Link)

[12] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. Zikopoulos, Understanding Big Data, New York, USA: McGraw-Hill, 2012.

[13] L. M. Ellram and W. L. Tate, "The use of secondary data in purchasing and supply management (P/SM) research," *Journal of Purchasing and Supply Management*, vol. 22, no. 4, pp. 250-254, Dec. 2016. Article (CrossRef Link)

[14] K. Coussement and D. Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction," *Information and Management*, vol. 45, no. 3, pp. 164-174, 2008. Article (CrossRef Link)

[15] M. A. Hearst, "Untangling Text Data Mining," in *Proc. of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 3-10, June 1999. Article (CrossRef Link)

[16] H. Hussein, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," *Computers in Human Behavior*, vol. 51, pp. 729-733, Oct. 2015. Article (CrossRef Link)
[17] J. Grimmer and B. M. Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis*, vol. 21, no. 3, pp. 267-297, 2013. Article (CrossRef Link)
[18] J. Ansell, T. Harrison, and T. Archibald, "Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis," *Marketing Intelligence and Planning*, vol. 25, no. 4, pp. 394-410, June 2007. Article (CrossRef Link)
[19] O. Malms and C. Schmitz, "Cross-Divisional Orientation: Antecedents and Effects on Cross-Selling Success," *Journal of Business-to-Business Marketing*, vol. 18, no. 3, pp. 253-275, Aug. 2011. Article (CrossRef Link)
[20] M. A. Jones, D. L. Mothersbaugh, and S. E. Beatty, "Why customers stay: measuring the underlying dimensions of services switching costs and managing their differential strategic outcomes," *Journal of Business Research*, vol. 55, no. 6, pp. 441-450, June 2002. Article (CrossRef Link)
[21] S. K. Saha, A. Aman, M. S. Hossain, A. Islam, and R. S. Rodela, "A Comparative Study On B2B Vs. B2C Based On Asia Pacific Region," *International Journal of Scientific and Technology Research*, vol. 3, no 9, pp. 294-298, 2014. Article (CrossRef Link)
[22] M. Subramani and E. Walden, "Economic Returns to Firms from Business-to Business Electronic Commerce Initiatives: An Empirical Examination, Association for Information Systems," in *Proc. of International Conference on Information Systems*, pp. 229-241, 2000. Article (CrossRef Link)
[23] A. S. Dick and K. Basu, "Customer Loyalty: Toward an Integrated Conceptual Framework," *Journal of the Academy of Marketing Science*, vol. 22, no. 2, pp. 99-113, 1994. Article (CrossRef Link)
[24] N. Behravan and M. SabbirRahman, "Customer Relationship Management Constructs under Social Networks towards Customers' Retention," *Australian Journal of Basic and Applied Sciences*, vol. 6, pp. 271-282, 2012. Article (CrossRef Link)
[25] S. N. Danesh, S. A. Nasab, and K. C. Ling, "The Study of Customer Satisfaction, Customer Trust and Switching Barriers on Customer Retention in Malaysia Hypermarkets," *International Journal of Business and Management*, vol. 7, no. 7, pp. 141-150, Apr. 2012. Article (CrossRef Link)
[26] C. Bonanni, J. Dermine, and L. H. Röller, "Some evidence on customer 'lock-in' in the French mutual funds industry," *Applied Economics Letters*, vol. 5, no. 5, pp. 275-279, 1998. Article (CrossRef Link)
[27] H. S. Yuan, Y. L. Wei, and X. G. Wang, "Maxent modeling for predicting the potential distribution of Sanghuang, an important group of medicinal fungi in China," *Fungal Ecology*, vol. 17, pp. 140-145, 2015. Article (CrossRef Link)
[28] A. P. Bradley, "The Use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, July 1997. Article (CrossRef Link)
[29] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk, "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves," *Surgery*, vol. 159, no. 6, pp. 1638-1645, Mar. 2016. Article (CrossRef Link)
[30] B. D. Arie, "About the relationship between ROC curves and Cohen's kappa," *Engineering Applications of Artificial Intelligence* 21, vol. 21, no. 6, pp. 874-882, Sep. 2008. Article (CrossRef Link)
[31] A. Cano, A. Zafra, and S. Ventura, "Weighted Data Gravitation Classification for Standard and Imbalanced Data," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1672-1687, Dec. 2013. Article (CrossRef Link)
[32] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276-282, Oct. 2012. Article (CrossRef Link)

[33] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513-523, 1988. Article (CrossRef Link)
[34] A. Aizawa, "An information-theoretic perspective of TF–IDF measures," *Information Processing and Management*, vol. 39, no. 1, pp. 45-65, Jan. 2003. Article (CrossRef Link)
[35] T. Xia and Y. Chai, "An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm," *Journal of Software*, vol. 6, no. 3, pp. 413-420, Mar. 2011. Article (CrossRef Link)
[36] L. Lopes, P. Fernandes, and R. Vieira, "Estimating term domain relevance through term frequency, disjoint corpora frequency-TF-DCF," *Knowledge-Based Systems*, vol. 97, pp. 237-249, Apr. 2016. Article (CrossRef Link)
[37] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 60, no. 5, pp. 493-502, Oct. 2004. Article (CrossRef Link)
[38] O. B. Petrina, V. Volokhova, S. E. Yalovitsyna, A. G. Varfolomeyev, and D. G. Korzun, "On Semantic Network Design for a Smart Museum of Everyday Life History," in *Proc. of the 20th Conference of FRUCT Association*, vol. 776, no. 20, pp. 676-680, 2017. Article (CrossRef Link)
[39] J. L. Myers and E. J. O'Brien, "Accessing the discourse representation during reading," *Discourse Processes*, vol. 26, no 2, pp. 131-157, Nov. 2009. Article (CrossRef Link)
[40] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and M. Jr. Wagner, "Word co-occurrence features for text classification," *Information Systems*, vol. 36, no. 5, pp. 843-858, July 2011. Article (CrossRef Link)
[41] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14-46, Nov. 2015. Article (CrossRef Link)
[42] Y. Hongliang, Z. H. Deng, and S. Li, "Identifying Sentiment Words Using an Optimization-based Model without Seed Words," in *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 855-859, Aug. 2013. Article (CrossRef Link)
[43] R. A. Stine, "Sentiment Analysis," *Annual Review of Statistics and its Application*, vol. 6, pp. 287-308, Mar. 2019. Article (CrossRef Link)
[44] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan, "Chest Pathology Detection Using Deep Learning With Non-Medical Training," in *Proc. of IEEE 12th International Symposium on Biomedical Imaging*, pp. 294-297, July 2015. Article (CrossRef Link)
[45] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998. Article (CrossRef Link)
[46] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1746-1751, 2014. Article (CrossRef Link)
[47] C. Li, L. Ji, and J. Yan, "Acronym Disambiguation Using Word Embedding," in *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 4178-4179, 2015. Article (CrossRef Link)
[48] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015. Article (CrossRef Link)
[49] L. Fan, "Revisit Fuzzy Neural Network: Demystifying Batch Normalization and ReLU with Generalized Hamming Network," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, Oct. 2017. Article (CrossRef Link)
[50] R. Mu and X. Zeng, "A Review of Deep Learning Research," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 4, pp. 1738-1764, 2019. Article (CrossRef Link)
[51] P. Vijayaraghavan, I. Sysoev, S. Vosoughi, and D. Roy, "Detecting Stance in Tweets Using Character and Word-Level CNNs," in *Proc. of the 10th International Workshop on Semantic Evaluation*, 2016. Article (CrossRef Link)

[52] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-Aware Neural Language Models," *Association for the Advancement of Artificial Intelligence*, pp. 2741-2749, 2015. Article (CrossRef Link)
[53] S. Rotchanakitumnuai and M. Speece, "Barriers to Internet banking adoption: a qualitative study among corporate customers in Thailand," *International Journal of Bank Marketing*, vol. 21, no. 6, pp. 312-323, 2003. Article (CrossRef Link)
[54] P. Thammasorn, L. W. A. Chaovlitwongse, L. Wootton, E. Ford, and M. Nyflot, "Deep convolutional Triplet network for quantitative medical image analysis with comparative case study of gamma image classification," in *Proc. of 2017 IEEE International Conference on Bioinformatics and Biomedicine*, 2017. Article (CrossRef Link)
[55] H. Bharadhwaj and S. Joshi, "Explanations for Temporal Recommendations," *Künstliche Intelligenz*, vol. 32, pp. 267-272, 2018. Article (CrossRef Link)
[56] X. Zhang, Z. Junbo, and Y. LeCun, "Character-level Convolutional Networks for Text Classification," in *Proc. of Advances in Neural Information Processing Systems*, 2015. Article (CrossRef Link)

Dr. Hoon Jung is a senior research fellow and general director at Hana Institute of Finance. He received a B.A. in Business Administration from Sung Kyun Kwan University and MBA from Seoul National University. He also received MBA from IE Business School and Ph.D. in Business Administration from Yonsei University. His research interests include machine learning, text mining, and database marketing.

Dr. Bong Gyou Lee is a Vice President/CIO and professor at Graduate School of Information in Yonsei University. He also has served as a director of Communications Policy Research Center since 2009. Dr. Lee received a B.A. from the Department of Economics at Yonsei University and he also received his M.S. and Ph.D. from Cornell University. He was a Commissioner of the Korea Communications Commission in 2007 and 2008.